Acta Cytologica

Nongynecologic Cytopathology

Acta Cytologica 2022;66:46–54
DOI: 10.1159/000519273

Received: June 15, 2021
Accepted: August 26, 2021
Published online: October 18, 2021

The Dynamics of Pathology Dataset Creation Using Urine Cytology as an Example

Ewen David McAlpine[a] Pamela M. Michelow[a] Turgay Celik[b]

[a]National Health Laboratory Service and Division of Anatomical Pathology, University of the Witwatersrand, Johannesburg, South Africa; [b]School of Electrical and Information Engineering and Wits Institute of Data Science, University of the Witwatersrand, Johannesburg, South Africa

Keywords
Digital pathology · Machine learning · Urine cytology · The Paris System

Abstract
Introduction: Dataset creation is one of the first tasks required for training AI algorithms but is underestimated in pathology. High-quality data are essential for training algorithms and data should be labelled accurately and include sufficient morphological diversity. The dynamics and challenges of labelling a urine cytology dataset using The Paris System (TPS) criteria are presented. ***Methods:*** 2,454 images were labelled by pathologist consensus via video conferencing over a 14-day period. During the labelling sessions, the dynamics of the labelling process were recorded. Quality assurance images were randomly selected from images labelled in previous sessions within this study and randomly distributed throughout new labelling sessions. To assess the effect of time on the labelling process, the labelled set of images was split into 2 groups according to the median relative label time and the time taken to label images and intersession agreement were assessed. ***Results:*** Labelling sessions ranged from 24 m 11 s to 41 m 06 s in length, with a median of 33 m 47 s. The majority of the 2,454 images were labelled as benign urothelial cells, with atypical and malignant urothelial cells more sparsely represented. The time taken to label individual images ranged from 1 s to 42 s with a median of 2.9 s. Labelling times differed significantly among categories, with the median label time for the atypical urothelial category being 7.2 s, followed by the malignant urothelial category at 3.8 s and the benign urothelial category at 2.9 s. The overall intersession agreement for quality assurance images was substantial. The level of agreement differed among classes of urothelial cells – benign and malignant urothelial cell classes showed almost perfect agreement and the atypical urothelial cell class showed moderate agreement. Image labelling times seemed to speed up, and there was no evidence of worsening of intersession agreement with session time. ***Discussion/Conclusion:*** Important aspects of pathology dataset creation are presented, illustrating the significant resources required for labelling a large dataset. We present evidence that the time taken to categorise urine cytology images varies by diagnosis/class. The known challenges relating to the reproducibility of the AUC (atypical) category in TPS when compared to the NHGUC (benign) or HGUC (malignant) categories is also confirmed.

Introduction

Dataset creation is one of the first tasks required for training supervised machine learning (ML) algorithms [1] and involves assigning a label (e.g., a diagnostic category) to a set of input variables (e.g., an image) – called annotation. Annotation can be burdensome, especially in a unique field such as cytopathology, as it needs to be un-

karger@karger.com
www.karger.com/acy

Correspondence to:
Ewen David McAlpine, ewen.mcalpine@wits.ac.za

Karger

dertaken by experts in the field [2, 3]. In fact, obtaining such annotated datasets are considered an underestimated hurdle for the development of ML in pathology in general [4] and is challenging, expensive, time-consuming, and considered tedious by some [5]. As artificial intelligence gains traction in pathology, pathologists will need to validate algorithm performance – even commercially procured algorithms – in their own laboratories which will also require knowledge of dataset creation techniques [1]. High-quality data are essential for training and evaluating supervised ML algorithms and these algorithms require data of sufficient quantity and quality to perform reliably in the real world [6]. Good quality datasets should be labelled as accurately as possible and be unbiased with respect to outliers [6]. In addition, pathology datasets should include sufficient diversity to account for variations in illumination, focus, and staining as well as differences in morphology [4]. It has also been suggested that, in the context of a field such as cytopathology, the behaviour of the expert annotators be monitored and that quality assurance (QA) measures are undertaken during the labelling process [4]. Guidelines aimed at improving dataset creation and ML model evaluation in pathology have been published by Marée [4]. These guidelines include suggestions to minimize bias in datasets created by technical and biological variability and important factors to consider during model evaluation and fine tuning. Pathologists, working in conjunction with ML practitioners, should keep these guidelines in mind when annotating pathology data. A brief summary of important factors to consider when creating pathology-specific datasets is provided in Table 1.

The goal of this article was to discuss the dynamics of labelling a small dataset of urine cytology images, highlight the challenges inherent in pathology dataset creation, and to provide brief guidance for dataset creation by pathologists. Specifically, we attempt to quantify the time taken to label images and whether labelling time differs by diagnostic category. In addition, we assess intra-observer variability in the labelling process and whether longer labelling sessions results poorer annotation performance. Previous literature, although not directly related to pathology, indicates that prolonged visual tasks may result in a decrease in the speed of information evaluation and decision-making [7], and the authors suggest that this is important to evaluate in a pathology context.

Presently, The Paris System (TPS) for Reporting Urine Cytology [8], introduced in 2016, attempts to standardize the reporting of urine cytology and to improve the reliable detection of high-grade urothelial carcinoma. Prior to the introduction of TPS, urine cytology was plagued by relatively poor intra- and interobserver variability [9]. Studies to assess intra- and interobserver variability conducted after the introduction of TPS have shown mixed results [10–12], but despite the persistence of challenges relating to intra- and interobserver variability in urine cytology categorization, TPS has led to measurable improvements in the specificity, positive predictive value, and diagnostic accuracy of urine cytology [13].

Furthermore, our data illustrates another important aspect of dataset creation that is pivotal to pathology – that of balanced datasets. This relates to the fact that each diagnostic category (or *class*) needs to be equally represented in the dataset [4, 5] to maximize the accuracy of an algorithm. In pathology, this is not always easy to achieve as rarer entities/diagnoses may be underrepresented in datasets and may require the use of additional techniques to address [4, 14].

Materials and Methods

A dataset of 2,454 512 × 512px images derived from digitized urine cytology slides was labelled. A total of 214 urine cytology slides were digitized using a Pannoramic 250 digital scanner (3DHISTECH Ltd., Budapest, Hungary) at ×400 magnification, resulting in 495,320 512 × 512px image patches. The slides were obtained from the archives of 2 independent laboratories, both located in Johannesburg, South Africa. Cases were retrieved retrospectively from June 2020. Cases of low-grade urothelial carcinoma were excluded. The 495,320 images were grouped into 1,000 clusters using K-means clustering with cosine similarity using deep features extracted by a pre-trained Resnet50 neural network. Initial screening of the clustering results selected 818 clusters for labelling and excluded 182 clusters that contained excessive obscuring inflammation or blood, artefact (including cracked coverslips) or only crystals. Examples of the types of images included and rejected after initial clustering are presented in Figure 1. The 3 images closest to the cluster centroid of each of the 818 clusters were manually labelled by 2 experienced pathologists by video conference using custom labelling software. Seven labelling sessions were conducted over a 14-day period. All sessions were held in the evenings starting approximately between 19:00 and 19:40 pm.

During the labelling process, the following variables were recorded: label time in milliseconds (defined as the time between displaying the image to the pathologists and assignment to a category), the relative time the image was labelled in minutes and seconds measured from the start of the labelling session, the assigned category (benign urothelial, atypical urothelial, malignant urothelial, and squamous cells or reject) and whether the image was a repeat, QA image.

TPS [8] for Reporting Urinary Cytology (2016) was used to classify urothelial cells, mimicking a real-world diagnostic setting without biopsy correlation. Histologic follow-up is not readily available in our local setting, and thus, we simulated our everyday diagnostic

Table 1. Important factors to consider when creating a cytology dataset for ML [3, 4, 6, 19]

Ensure that a dataset accounts for the variability encountered in the real world
Accounting for technical differences – including differences in staining, technical staff preparing the samples, focus, image acquisition methodology/settings (including using different slide scanners) and examples from different laboratories
Accounting for the morphologic variability in pathologic material
Magnification used in the real world
Include an "other/reject" category to account for non-cellular objects and artefact
Ensure that the dataset is of sufficient quality
All objects/classes should be as equally represented in the dataset as possible
Datasets should lack outliers or missing features
Annotator behaviour should be monitored during the annotation process to minimize labelling bias and annotator fatigue
Ensure appropriate QA in the annotation process
Pathology datasets should be annotated by domain experts (e.g., pathologists, cytologists, and technologists)
All annotators should be familiar with, and use the same criteria for classification (e.g., accepted cytology classification systems, such as TPS)
Inter- and intra-observer variability/agreement should be monitored during the annotation process
Ensure algorithms are evaluated appropriately
A sufficient number of examples should be labelled to allow for a dataset to be split into 3 parts – a training set, a validation set, and a test set
The training set is used to train the algorithm, the validation set is used to fine tune the model's parameters, and the test set is a completely unseen set used to assess the model's performance on real-world data
The evaluation criteria used can include metrics commonly used in ML (e.g., accuracy, specificity, and sensitivity) and outcomes used by pathologists in the real world
Publish details about data acquisition and annotation processes
When publishing a dataset, release details pertaining to the data acquisition and annotation process to allow those who make use of the dataset to investigate potential sources of bias and poor performance

ML, machine learning; TPS, The Paris System.

environment where cytological criteria, such as TPS are applied without histologic correlation. Images containing benign urothelial cells were assigned to the *Benign* class. Images containing cells with mild atypia – defined as a nuclear to cytoplasmic (N/C) ratio exceeding 0.5 with 1 or more of the following additional features: mild nuclear hyperchromasia, irregular nuclear contours, or irregular coarse chromatin – were assigned to the *Atypical* class. When severe atypia was present – as defined by an N/C ratio >0.7 together with hyperchromasia and 1 or more of the following additional features: coarse chromatin or irregular nuclear membranes – images were assigned to the *Malignant* class. The current Paris System separates suspicious for high-grade urothelial carcinoma and high-grade urothelial carcinoma (HGUC) depending on the number of severely atypical cells. In the present study, these 2 categories were combined into a single *Malignant* class as quantification of cells is difficult on 512 × 512px static images.

For each of the labelling sessions, excluding the first, QA images were randomly selected from urothelial cells labelled in the previous labelling session and then randomly distributed throughout the new labelling session. Images labelled as squamous cells and those rejected were excluded from the QA process.

To assess the effect of time on the labelling process, the labelled set of images was split into 2 groups – images labelled up to and including the median of the time the images were labelled relative to the start (termed the relative label time) of the labelling session and images labelled after this value. Specifically, the label time and intersession agreement, measured by the Cohen κ coefficient, were assessed.

Statistical Methods

Categorical data has been summarized using frequencies and percentages. Non-parametric methods were used to analyse numeric variables due to the lack of normal distribution and the presence of outliers. The differences in labelling time among classes were assessed using a Kruskal-Wallis test. To assess the significance of the difference in labelling time among classes, pairwise 2-tailed Kolmogorov-Smirnov tests and a Dunn test with Bonferroni adjustment were performed. A Mann-Whitney test was used to assess the difference in labelling time per image in the 2 groups of images separated according to relative label time. A p value of 0.05 was used to determine statistical significance. Intersession/intra-observer agreement of urothelial categories was assessed by way of a κ coefficient. κ values between 0.0 and 0.20 indicated slight agreement, between 0.21 and 0.40 fair agreement, 0.41–0.60 moderate agreement, 0.61–0.80 substantial agreement, and 0.81–1.00 almost perfect agreement [15].

Software Used

Whole slide image processing and graph creation were performed in Python™ 3 (https://www.python.org/). Feature extraction and clustering were performed using Matlab™ R2021a (http://www.mathworks.com/). The custom image labelling software was written in C# (Microsoft® Corporation), and Zoom™ (http://zoom.us) was used for teleconferencing. Statistical analysis was conducted in the statistical analysis software R [16].

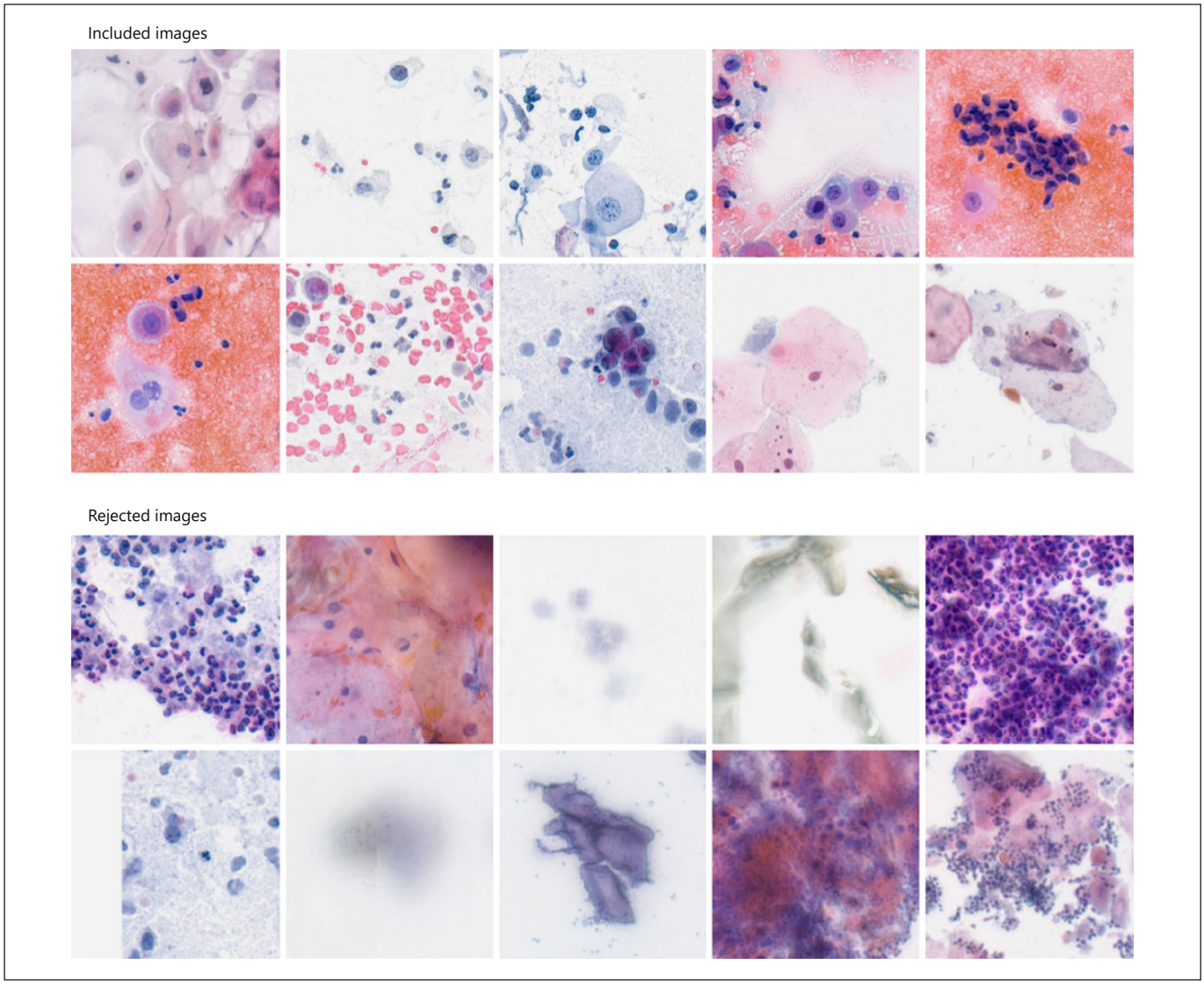

Fig. 1. Examples of the types of images included and rejected after initial clustering.

Results

Between 200 and 445 images were labelled during sessions that ranged from 24 m 11 s to 41 m 06 s in length. The median length of the labelling session was 33 m 47 s (interquartile range = 07 m 38 s).

Of the 2,454 unique images, 1,153 (46.99%) were labelled as benign urothelial cells, 187 (7.62%) as atypical urothelial cells, and 175 (7.13%) as malignant urothelial cells. A further 403 (16.42%) images were labelled as squamous cells and 536 (21.84%) were rejected as they contained excessive obscuring inflammation, were out of focus or did not contain urothelial or squamous cells. Table 2 summarizes the

Table 2. Summary of the annotated urine cytology dataset

	New images, frequency (%)	QA images, frequency (%)
B	1,153 (46.99)	188 (77.05)
AI	187 (7.62)	29 (11.89)
M*	175 (7.13)	27 (11.07)
S	403 (16.42)	–
R	536 (21.84)	–
Total	**2,454**	**244**

QA, quality assurance; SHGUC, suspicious for high-grade urothelial carcinoma; HGUC, high-grade urothelial carcinoma; B, benign; A, atypical; M, malignant; S, squamous; R, reject; TPS, The Paris System. *Includes TPS categories SHGUC and HGUC.

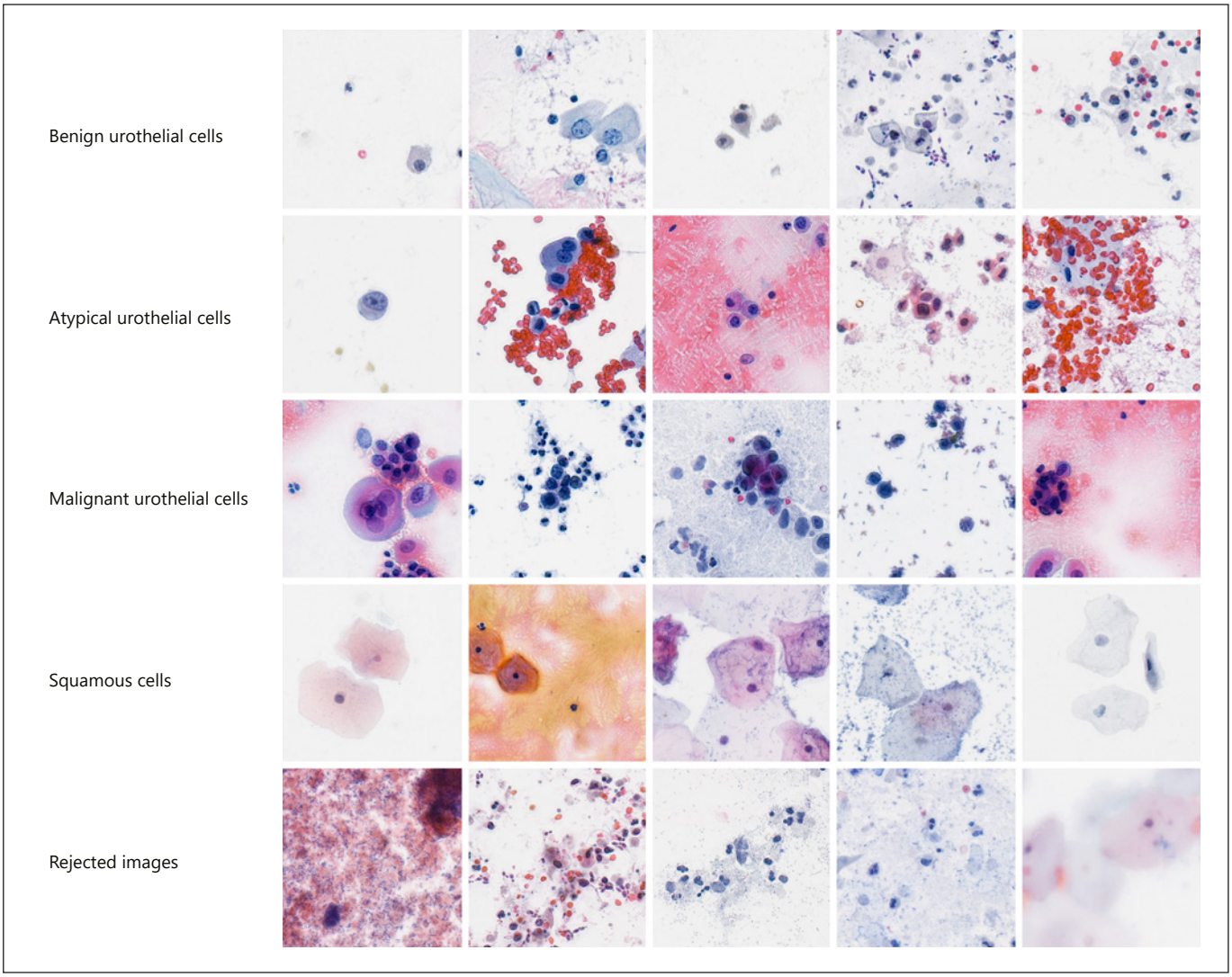

Fig. 2. Examples of images included in the dataset.

breakdown of the dataset and Figure 2 demonstrates examples of images in the dataset.

The time taken to label individual images ranged from 1 s to 42 s with a median of 2.9 s and an interquartile range of 1.9 s. The median time taken to assign an image to the atypical urothelial category was the slowest at 7.2 s, followed by the malignant urothelial category at 3.8 s. Assigning images to the benign urothelial category had the shortest median time of the 3 urothelial categories at 2.9 s. Labelling an image as squamous cells or rejecting an image was faster than assigning an image to a urothelial category, with median times of 2.1 and 2.7 s, respectively. The differences in the time take to label images among categories were statistically significant ($p < 0.001$, Kruskal-Wallis test). Furthermore, the significant difference in labelling time among categories was present across all pairwise combinations of the categories ($p < 0.01$ for all categories by both Dunn post hoc test and Kolmogorov-Smirnov tests). Figure 3 illustrates the labelling time of image patches by category.

The overall level of agreement for all 244 QA images was substantial ($\kappa = 0.73$). The level of agreement differed among classes of urothelial cells, with the benign and malignant urothelial cell classes showing almost perfect agreement (κ scores of 0.82 and 0.81, respectively) while the atypical urothelial cell class showed moderate agreement ($\kappa = 0.49$). The results of the QA process conducted during the annotation process is

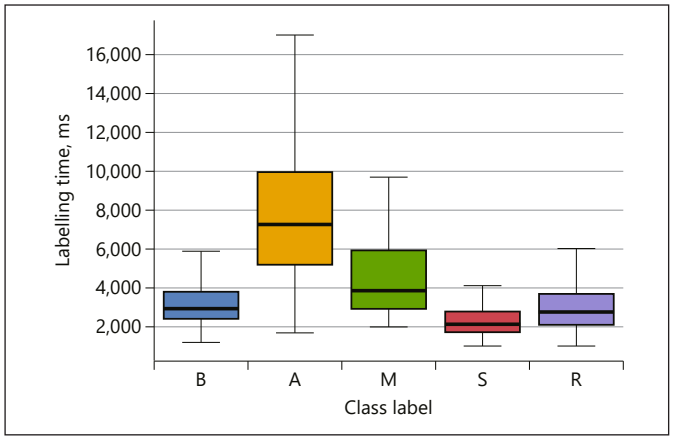

Fig. 3. Image labelling times by diagnostic category. B, benign; A, atypical; M, malignant; S, squamous; R, reject.

Table 3. Summary of the QA dataset and associated level of agreement

Urothelial cell category	Frequency (%)	κ score	Level of agreement
B	188 (77.05)	0.82	Almost perfect agreement
A	29 (11.89)	0.49	Moderate agreement
M	27 (11.07)	0.81	Almost perfect agreement
Overall		0.73	Substantial agreement

QA, quality assurance; B, benign; A, atypical; M, malignant.

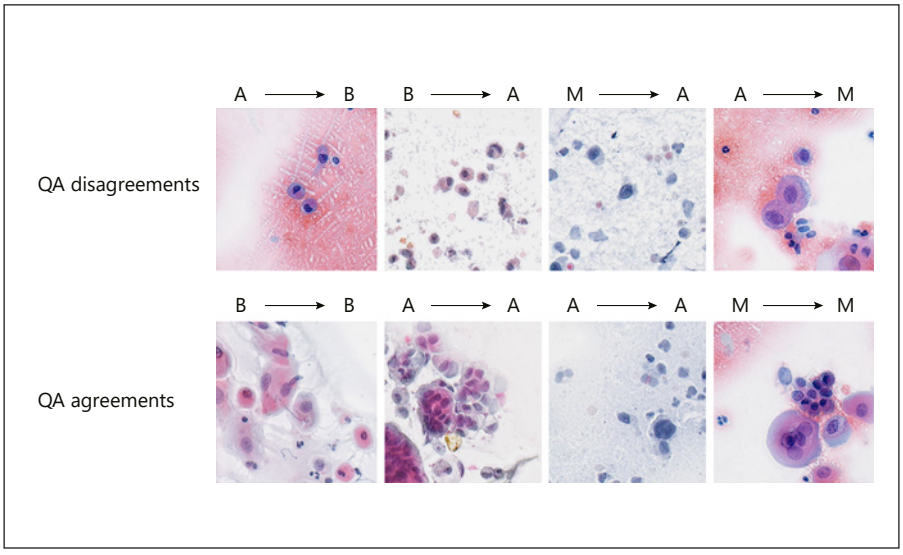

Fig. 4. Examples of QA images and the class label assigned to each image in the first and second labelling sessions. B, benign; A, atypical; M, malignant; QA, quality assurance.

presented in Table 3 and examples of QA images are shown in Figure 4.

To determine if prolonged labelling sessions caused fatigue or if there was a difference in label time in the first or latter part of the labelling session, a median relative label time was used as a cut off to evaluate the first and seconds halves of the images labelled. The median relative label time was 18 m 23 s and this value was used as the cut off to split the dataset into 2 groups – the first comprising images labelled up to and including this value and the second, images labelled after this value. The time taken to label each image in the first group ranged from 1 s to 24.9 s, with a median of 3.1 s and an interquartile range of 2.1 s. The time taken to label each image in the second group ranged from 1 s to 42 s, with a median of 2.8 s and an interquartile range of 1.6 s. The difference in labelling time per image in each group was statistically significant ($p < 0.001$, Mann-Whitney test). Figure 5 demonstrates image labelling time in these 2 groups.

The level of agreement in the QA set in these 2 groups was assessed by measuring the κ score in the following groups: both images labelled before 18 m 23 s, 1 image labelled after 18 m 23 s, and both images labelled after 18 m 23 s. The first set (61 images) showed moderate agreement ($\kappa = 0.56$), the second group (118 images) showed almost perfect agreement ($\kappa = 0.81$), and the third group (65 images) showed substantial agreement ($\kappa = 0.78$), indicating that the level of agreement did not get worse as the relative label time increased. These data are summarized in Table 4.

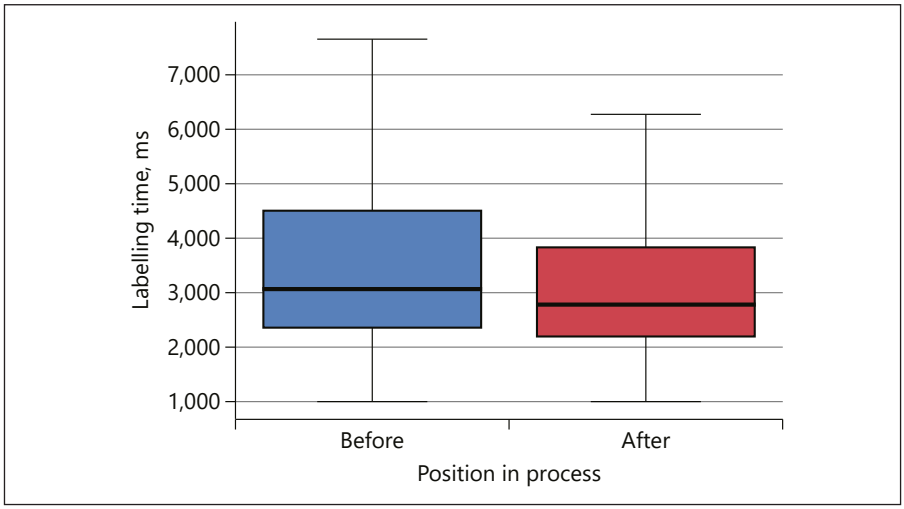

Fig. 5. Image labelling time for images before and after the median relative label time.

Table 4. Summary of the level of agreement between urothelial categories before and after the median relative label time

Category	Both images labelled before median relative label time	One image labelled before median relative label time	Both images labelled after median relative label time
Images, *n*	61	118	65
κ score	0.56	0.81	0.78
Level of agreement	Moderate agreement	Almost perfect agreement	Substantial agreement

Discussion/Conclusion

Dataset creation is a fundamental component of implementing ML in pathology, including in the training and validation of algorithms in laboratories, but limited pathology literature regarding the dynamics of the dataset creation process exists. In the current study, important insights were gleaned by analysing the dynamics of creating a modest urine cytology dataset. Dataset creation, although an important preliminary step in implementing ML algorithms in cytology, adds to the workload of pathologists and is likely to become a part of routine practice in the future. Previously published data indicate that dataset labelling is a time-consuming process, taking up to 120–130 h of expert time to label 50 image patches for nuclear segmentation algorithm training [17]. Datasets can be annotated to differing degrees of detail [1]. Image segmentation, representing strongly labelled data where each pixel in an image is assigned a categorical label, requires a significant amount of time to complete. Weaker annotation, such as assigning a label on an image-wise basis, as was performed in this study, will require less pathologist/cytotechnologist time. The median time taken for assigning a label to the images in our study was 2.9 s. Of note, however, is that image labelling time took up to 42 s for more difficult images. Interestingly, the time taken to assign images to different categories differed significantly. Assigning an image to urothelial category took longer than assigning an image to the squamous cell category or rejecting the image outright. Amongst the urothelial categories, assigning an image to the atypical group took almost twice as long as the malignant group and almost two-and-a-half times as long as the benign group. Extrapolating from this limited dataset, it is estimated that labelling the entire 495,320 images obtained from the WSI of the 214 cases used in this work would take approximately 400 h of pathologist time. Even if a pathologist could label continuously for an hour, longer than any of the sessions undertaken in the present study, this would mean 400 separate labelling sessions. Notably, this excludes QA images in the process which would necessarily lengthen the labelling process and incorporating QA procedures in the dataset creation process is considered best practice [4].

As already stated, inter- and intra-observer variability in urine cytology remains a challenge despite the introduction of TPS. In the present study, the authors randomly introduced previously labelled urothelial images into the subsequent labelling sessions to assess intersession

agreement. A total of 244 images initially placed into urothelial category were relabelled blindly. The overall agreement, as measured by a κ statistic, was substantial although, as with labelling time, differences amongst diagnostic categories were evident. While the benign and malignant classes showed almost perfect intersession agreement, the atypical class showed only moderate agreement. The difficulty in the reproducibility of the AUC category is well documented in the literature, despite specific TPS criteria. The Paris Interobserver Reproducibility Study (PIRST) [11], an online survey published in 2018 which aimed to assess diagnostic agreement amongst participants and the TPS author consensus, showed similar findings to the present study, with the most disagreement in the atypical category and the most agreement in both NHGUC (benign) and HGUC (malignant) categories. In a 2017 study by Long et al. [10], the authors showed adequate interobserver reproducibility for NHGUC while deeming the reproducibility of the other TPS categories unacceptable. In contrast to the Paris Interobserver Reproducibility Study (PIRST) [11], Long et al. found poor agreement amongst pathologists when diagnosing HGUC. A more recent study by Wang et al. [12], again conducted as an online survey, showed poor interobserver concordance of both diagnosis and cytologic criteria in urine cytology. As TPS places a strong emphasis on (N/C ratio for categorization of urothelial cells, Long et al. [10], investigated the accuracy and reproducibility of assessing N/C ratios amongst 6 pathologists. The authors found fair correlation between N/C estimation by pathologists and actual N/C ratio determined by image analysis. Additionally, moderate interrater agreement of N/C ratio was noted by Long et al. [10]. The ability of pathologists to accurately estimate N/C has been questioned by Zhang et al. [18], who found that practitioners tend to overestimate N/C; however, they did note that these "morphologists" (pathologists, technologist, and pathology residents) were significantly more accurate than non-pathology trained individuals. The predominant consensus from available literature suggests that while TPS has led to improvements in the specificity, positive predictive value, and diagnostic accuracy of urine cytology [13], this improvement is most pronounced in the NHGUC (benign) and HGUC (malignant) categories [12], and that the reproducibility of the AUC (atypical) category remains challenging [10–12], a fact supported by the present study.

Visual fatigue is an important factor that may affect the accuracy of pathology dataset labelling. Gou et al. showed that prolonged visual attention tasks lead to evidence of observer fatigue as measured by both subjective (fatigue rating) and objective (e.g., reaction times and accuracy rates) measures. The authors measured fatigue levels of participants in the time period 0–25 min and again in 36–60 min. While our labelling sessions were relatively short ranging from just over 24 min to around 41 min, with a median of 33 m 47 s, we attempted to assess the effect of session length on the labelling process by measuring image labelling time and intersession agreement in the first and second halves of the labelling sessions (as defined by the median *relative label time*). Our median *relative label time* was 18 m 23 s, and our data show no evidence of a decrease in either reaction time (measured by image labelling time) or accuracy (measured by intersession agreement in QA images). In fact, our data suggest a significant decrease in image labelling time in the second half of the labelling session. This trend of faster labelling times and no evidence of worsening of intersession agreement are not; however, expected to persist with increasing labelling session time. Further research into the optimal length of time for labelling sessions and assessing visual fatigue in pathology dataset creation is warranted.

Lastly, our dataset illustrates a potential problem for creating datasets for training ML algorithms in pathology. Datasets used to train classification algorithms should contain roughly equal numbers of examples of each diagnostic class the algorithm will be trained to identify. This is referred to as a balanced dataset [4, 5]. Our dataset is an example of an unbalanced dataset with over 75% of urothelial images being benign and just over 10% of urothelial images being atypical or malignant. Pathology datasets may be prone to being unbalanced because rare diseases or examples will be difficult to locate, identify, and label for inclusion in training sets. Specifically, in urine cytology, benign urothelial cells are more abundantly represented, creating a potential source of bias in urine cytology dataset creation.

In summary, this study contributes important findings and practical guidance relating to the process of pathology dataset creation. Specifically, proposing and implementing techniques specifically related to pathology datasets, quantifying the time taken to label a modest dataset in ML terms, and illustrating the enormous resources required for labelling a large dataset by extrapolation of these measurements. Additionally, we present evidence that the time taken to categorize urine cytology images varies by diagnosis/class and that labelling an image as atypical takes more time than assigning an image to a benign or malignant class. The known challenges relating to the reproducibility of the AUC category in TPS when compared to the NHGUC (benign) or HGUC (malignant) categories is also confirmed by our data.

Acknowledgment

The authors acknowledge the National Health Laboratory Service, Lancet Laboratories, and 3F Scientific for their assistance with providing research material and digitising the slides in this research, respectively; and Mr Eric Liebenberg for maintaining the computer equipment used in this research.

Statement of Ethics

This study involved the use of anonymized, archived cytology slides. Ethical clearance for this study was granted by the Human Research Ethics Committee (Medical) of the University of the Witwatersrand (Certificate No. M190604).

Conflict of Interest Statement

The authors have no conflicts of interest to declare.

Funding Sources

This work has been partially funded by University of the Witwatersrand, Faculty of Health Sciences Research Equipment grants.

Author Contributions

All 3 authors contributed to the conceptualization of this study, the analysis of the results, and the writing of the manuscript. The algorithms and custom software used in this project were written by E.M. and T.C. Image labelling was performed by E.M. and P.M.

Data Availability Statement

Due to constraints in the ethical approval and material exchange agreements with laboratories, only anonymized, summarized data can be made available by the authors. Further enquiries can be directed to the corresponding author.

References

1 McAlpine ED, Michelow P. The cytopathologist's role in developing and evaluating artificial intelligence in cytopathology practice. Cytopathology. 2020;31(5):385–92.
2 Peikari M, Salama S, Nofech-Mozes S, Martel AL. A cluster-then-label semi-supervised learning approach for pathology image classification. Sci Rep. 2018;8(1):7193–13.
3 Tizhoosh HR, Pantanowitz L. Artificial intelligence and digital pathology: challenges and opportunities. J Pathol Inform. 2018;9(1):38.
4 Marée R. The need for careful data collection for pattern recognition in digital pathology. J Pathol Inform. 2017 [cited 2018 Dec 2];8:19. Available from: http://www.jpathinformatics.org/text.asp?2017/8/1/19/204200.
5 Abels E, Pantanowitz L, Aeffner F, Zarella MD, vd Laak J, Bui MM, et al. Computational pathology definitions, best practices, and recommendations for regulatory guidance: a white paper from the Digital Pathology Association. J Pathol. 2019;249(3):286–94.
6 Géron A. Hands-on machine learning with scikit-learn, keras & tensorFlow. 2nd ed. O'Reilly Media Inc.; 2019. p. 23–32.
7 Guo Z, Chen R, Zhang K, Pan Y, Wu J. The impairing effect of mental fatigue on visual sustained attention under monotonous multi-object visual attention task in long durations: an event-related potential based study. PLoS One. 2016;11(9):1–13.
8 Barkan GA, Wojcik EM, Nayar R, Savic-Prince S, Quek ML, Kurtycz DF, et al. The Paris System for reporting urinary cytology: the quest to develop a standardized terminology. Adv Anat Pathol. 2016 [cited 2018 Dec 223(3):193–201.
9 Reid MD, Osunkoya AO, Siddiqui MT, Looney SW. Accuracy of grading of urothelial carcinoma on urine cytology: an analysis of interobserver and intraobserver agreement. Int J Clin Exp Pathol. 2012;5(9):882–91.
10 Long T, Layfield LJ, Esebua M, Frazier SR, Giorgadze DT, Schmidt RL. Interobserver reproducibility of the Paris system for reporting urinary cytology. Cytojournal. 2017;14(1):17.
11 Kurtycz DFI, Barkan GA, Pavelec DM, Rosenthal DL, Wojcik EM, VandenBussche CJ, et al. Paris Interobserver Reproducibility Study (PIRST). J Am Soc Cytopathol. 2018;7(4):174–84.
12 Wang YH, Hang JF, Wen CH, Liao KC, Lee WY, Lai CR. Diagnostic agreement for high-grade urothelial cell carcinoma in atypical urine cytology: a nationwide survey reveals a tendency for overestimation in specimens with an N/C ratio approaching 0.5. Cancers. 2020;12(2):272.
13 Stanzione N, Ahmed T, Fung PC, Cai D, Lu DY, Sumida LC, et al. The continual impact of the Paris System on urine cytology, a 3-year experience. Cytopathology. 2020;31(1):35–40.
14 Wei J, Suriawinata A, Vaickus L, Ren B, Liu X, Wei J, et al. Generative image translation for data augmentation in colorectal histopathology images. Proc Mach Learn Res. 2019;116:10–24.
15 Cuff J, Higgins JP. Statistical analysis of surgical pathology data using the R program. Adv Anat Pathol. 2012 19:131–9.
16 R Core Team. R: a language and environment for statistical computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2017. Available from: https://www.r-project.org/.
17 Hou L, Agarwal A, Samaras Di, Kurc TM, Gupta RR, Saltz JH. Robust histopathology image analysis: to label or to synthesize? Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit. 2019 Jun;2019:8525–34.
18 Zhang ML, Guo AX, VandenBussche CJ. Morphologists overestimate the nuclear-to-cytoplasmic ratio. Cancer Cytopathol. 2016;124(9):669–77.
19 Chollet F. Deep learning with python. 1st ed. Greenwich, CT, USA: Manning Publications Co.; 2017.